



**HAL**  
open science

## Universality of the DNA methylation codes in Eucaryotes

Benoît Aliaga, Ingo Bulla, Gabriel Mouahid, David Duval, Christoph Grunau

► **To cite this version:**

Benoît Aliaga, Ingo Bulla, Gabriel Mouahid, David Duval, Christoph Grunau. Universality of the DNA methylation codes in Eucaryotes. *Scientific Reports*, 2019, 9, pp.173. 10.1038/s41598-018-37407-8 . hal-01905101

**HAL Id: hal-01905101**

**<https://hal.science/hal-01905101>**

Submitted on 25 Oct 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 Universality of the DNA methylation codes in Eucaryotes

2 Benoît Aliaga<sup>1</sup>, Ingo Bulla<sup>1,2,3</sup>, Gabriel Mouahid<sup>1</sup>, David Duval<sup>1</sup> and Christoph  
3 Grunau<sup>1\*</sup>

4  
5 (1) Univ. Perpignan Via Domitia, IHPE UMR 5244, CNRS, IFREMER, Univ.  
6 Montpellier, F-66860 Perpignan, France

7 (2) Institute for Mathematics and Informatics, University of Greifswald,  
8 Greifswald, Germany

9 (3) Department of Computer Science, ETH Zürich, Zürich, Switzerland

10  
11 \* Correspondence to [christoph.grunau@univ-perp.fr](mailto:christoph.grunau@univ-perp.fr)  
12

## 13 **Abstract:**

14 Genetics and epigenetics are tightly linked heritable information classes. Question arises if  
15 epigenetics provides just a set of environment dependent instructions, or whether it is integral part  
16 of an inheritance system. We argued that in the latter case the epigenetic code should share the  
17 universality quality of the genetic code. We focused on DNA methylation. Since availability of  
18 DNA methylation data is biased towards model organisms we developed a method that uses kernel  
19 density estimations of CpG observed/expected ratios to infer DNA methylation types in any  
20 genome. We show here that our method allows for robust prediction of mosaic and full gene body  
21 methylation with a PPV of 1 and 0.87, respectively. We used this prediction to complement  
22 experimental data, and applied hierarchical clustering to identify methylation types in ~150  
23 eucaryotic species covering different body plans, reproduction types and living conditions. Our  
24 analysis indicates that there are only four gene body methylation types. These types do not follow  
25 phylogeny (i.e. phylogenetically distant clades can have identical methylation types) but they are  
26 consistent within clades. We conclude that the gene body DNA methylation codes have universality  
27 similar to the universality of the genetic code and should consequently be considered as part of the  
28 inheritance system.

29  
30  
31  
32  
33  
34

## 1 **Introduction**

2 Living organisms are biological systems in which the complex interaction between different  
3 elements such as the nuclear genotype and epigenotype factors and the environment brings about a  
4 phenotype that develops and evolves over time<sup>1,2</sup>. For a complete understanding and potential  
5 control of biological processes such as development and evolution, it is therefore necessary to  
6 understand as many elements of biological systems as possible. In the present work, we focus on  
7 the epigenotype unit that we operationally define as any modification of the chromatin-DNA  
8 complex that has an impact on the expression and function of genes<sup>3</sup>. Epigenetic information can be  
9 stored in a multitude of bearers such as histone modifications, non-coding RNA, the topology of the  
10 nucleus, and methylation of DNA. DNA methylation has been one of the most studied epigenetic  
11 marks since its discovery in 1948<sup>4</sup>. Methylation occurs at positions 4 and 5 of the pyrimidine ring of  
12 cytosine forming either 4-methyl-cytosine (4mC) or 5-methyl-cytosine (5mC), or at position 6 of  
13 the purine ring in 6-methyl-adenine (6mA). 6mA and 4mC were believed to occur only in bacteria  
14 but recent advances in sequencing technology made it possible to detect them also in eukaryotic  
15 species. A specific database (MethSMRT) was dedicated to these modifications<sup>5</sup>, and the available  
16 experimental data were used to train an algorithm to predict the occurrence of 4mC<sup>6</sup> in DNA based  
17 on sequence features. We will focus here on 5mC and to facilitate the readability use the term *DNA*  
18 *methylation* for this purpose.

19 In most eukaryotes, 5mC is overrepresented or even restricted to the dinucleotide CpG context,  
20 where ‘p’ stands for the phosphodiester linkage between the cytosine (C) and the guanine (G). In  
21 plants, the 5mC can occur in other contexts such as CpHpG or CpHpH, where ‘H’ stands for A, C  
22 or T (reviewed in Vanyushin<sup>7</sup>). In contrast, in certain molds, methylation occurs preferentially  
23 (>60%) in CpAs<sup>8</sup>. DNA methylation is catalyzed by a family of enzymes called DNA  
24 methyltransferase (DNMT) composed of 3 canonical members (DNMT 1, 2 and 3)<sup>9</sup>. After  
25 replication, 5mC will be maintained by the activity of DNMT1, which has a high affinity to hemi-  
26 methylated DNA, and that methylates immediately after replication the newly synthesized strand,  
27 reproducing methylation patterns in CpG dinucleotide with a fidelity of roughly 99.9%<sup>10</sup> thus  
28 allowing for mitotic heritability of DNA methylation patterns. The role of DNMT2 is controversial  
29 because it has little DNA methylation activity<sup>11</sup> but is able to methylate cytosine 38 in the anticodon  
30 loop of aspartic acid transfer RNA<sup>12</sup> and some authors propose therefore to replace DNMT2 by  
31 tRNA (Cytosine(38)-C(5))-Methyltransferase TRDMT1<sup>13</sup>. There are species, such as the model  
32 organism *Drosophila melanogaster*, that have only DNMT2 and do not possess 5-methyl-cytosine  
33 in their genome, or DNA methylations is so low that it is very difficult to detect<sup>11,14,15</sup>. These  
34 enzymes have distinct roles due to the presence of different domain structures. DNA methylation is

1 established by DNMT3 that can methylate the two strands of the DNA *i.e.* has a *de novo*  
2 methylation function<sup>16</sup>. Various DNA methylation contexts are found across the plant and animal  
3 kingdoms. There are species where 5mC is present all over the genome (global methylation) while  
4 others can be entirely devoid of methylation. In species with global DNA methylation, only small  
5 regions, among them promoters and other regulatory elements, are methylation-free<sup>17</sup>. If 5mC  
6 occurs in the promoters of vertebrates, it has a repressive action on the gene transcription<sup>18</sup>.  
7 Invertebrates often have a mosaic-type methylation pattern with high methylation in almost all CpG  
8 in large blocks of genomic DNA interspersed with almost entirely unmethylated blocks. Changes in  
9 DNA methylation occur during organ regeneration<sup>19</sup>, aging<sup>20,21</sup>, in response to bacterial infection<sup>22</sup>,  
10 as well as flowering time and root length in *Arabidopsis thaliana*<sup>23</sup>, just to name a few examples.  
11 DNA methylation was therefore proposed as a language in which environment and genome talk to  
12 each other. Other authors have seen DNA methylation primarily as a genomic defense system  
13 against parasitic genomic elements<sup>17</sup>.

14 Given the apparent heterogeneity of DNA methylation patterns and the multitude of biological  
15 processes involved it was suggested that DNA methylation evolved in every phylogenetic clade  
16 towards a specific role in controlling gene expression. Essentially, the question is whether or not  
17 there is universality in the DNA methylation epigenetic code, conceptually similar to the  
18 universality of the genetic code, or if DNA methylation is non-universal and specific to every  
19 evolutionary unit. We wished to address this question through the analysis of the evolution of DNA  
20 methylation. Here, an obstacle is that a comprehensive analysis of DNA methylation patterns in a  
21 wide range of different species is missing. There are currently methods available that allow, in  
22 principle, for determining genome-wide DNA methylation patterns (“methylomes”) at a single base  
23 resolution. Since DNA methylation patterns can be different in different organisms of a species or  
24 even tissues of an individual, for a given species several methylomes can exist. A review of the  
25 available data in different databases and in the literature showed that there is a strong bias towards  
26 model organisms: there are at least 300 methylomes available for human, mouse and the model  
27 plant *A. thaliana*, but only 63 for a total of 16 other species  
28 (<http://smithlabresearch.org/software/methbase/> and Céline Cosseau, pers. communication). As a  
29 consequence, global conclusions about the function and importance of DNA methylation are  
30 actually based on a very limited and biased amount of data. For this reason, it remains challenging  
31 to derive the general rules (if any) that govern DNA methylation in the different branches of the  
32 “tree of life”. A potential solution to the caveat that experimental “wet bench” data is missing is to  
33 infer DNA methylation indirectly with computational method<sup>24,25</sup>. The basis for this is that  
34 methylated CpG sites mutate relatively frequently compared to the other dinucleotides over  
35 evolutionary time<sup>26</sup>. If a cytosine is deaminated, a deoxy-uracil will form, which is not stable in

1 DNA: it will rapidly be excised by uracil glycosylase and replaced by cytosine. In contrast, 5mC  
2 deamination generates thymine, which is less efficiently processed by the DNA repair machinery.  
3 Despite the existence of a specific repair mechanism that restores G/C mismatch, the mutation rate  
4 from 5mC to T is therefore 10-fold to 50-fold higher than other transitions, depending on local GC  
5 content<sup>27</sup>. For Humans, it was estimated that within 20 years, 0.17% of 5mC in the genome were  
6 converted into thymine<sup>28</sup>. If 5mC occurs predominantly in CpG pairs, the above-mentioned  
7 mechanism will increase the mutation rate from CpG to TpG or CpA and induce an  
8 underrepresentation of CpG<sup>29</sup>. Therefore, CpG observed/expected ratio (CpG o/e) in gene bodies  
9 can be used to predict if a species' DNA is methylated in gene bodies or not<sup>26,30,31</sup>. In other words,  
10 gene bodies of a species are not methylated when the CpG o/e ratio is in average close to 1, and  
11 methylated for an average ratio far below 1. Low CpG o/e is not a condition for methylation but a  
12 consequence of it. It is important to note that this is a species-level prediction that uses methylation  
13 signatures that pass through the germ-line and need several generations of mutation accumulation to  
14 be detectable. It cannot be used to predict methylation changes of individual genes within shorter  
15 periods.

16 These predictions were tested in at least 13 studies comparing CpGo/e to methylation levels  
17 obtained with various methods (Table 1 and Supplementary file 1). All came to the conclusion that  
18 CpGo/e ratios correlate well (inversely) with methylation levels when species were compared.  
19 Nevertheless, there remain technical challenges. For instant, in the past, prediction of *in silico* DNA  
20 methylation based on Gaussian distributions, that are relatively straightforward to implement, were  
21 used to describe the frequency distribution of CpGo/e ratios<sup>32-35</sup>. But in many species, frequency  
22 distributions of CpGo/e ratios are complex or skewed and Gaussian distribution is not suitable. In  
23 our hands, only for 58 out of 83 cases (65%) Gaussian mixtures allowed for description of the  
24 distribution<sup>36</sup>. These values are comparable to what was found by Bewick and colleagues who used  
25 CpG o/e ratios in transcriptomes of 124 species of which only 50 (40.32%) were described correctly  
26 with Gaussian mixtures<sup>37</sup>. We have also tested non-Gaussian distributions and the results were even  
27 less conclusive than Gaussian distributions: out of 83 only 41% delivered an exploitable result.  
28 Therefore, we have developed a new tool, called Notos, to identify DNA methylation signatures  
29 within CpGo/e ratios based on kernel density estimations<sup>36</sup>. This novel algorithm delivers robust  
30 descriptions of frequency distributions of CpGo/e ratios for up to 172,000 input sequences.

31 Here, we have applied this software to predict DNA methylation with CpGo/e ratios in a total of  
32 634 species and to use the results in combination with publicly available experimental data to infer  
33 evolution of DNA methylation over the eukaryotic tree of life. We applied Notos on coding  
34 sequences coming from three databases (dbEST, CleanEST, and CDS/cDNA). Our results show  
35 clearly (i) that DNA methylation prediction by CpGo/e ratio is robust, (ii) that only four types of

1 DNA methylation can be identified in all species despite their wide range of genome sizes,  
2 environments, body plans, reproduction types etc., and (iii) that DNA methylation types does not  
3 follow phylogeny but is consistent within clades suggesting evolutionary constraints. Taken  
4 together our analysis delivers arguments for the idea of the universality of the role of DNA  
5 methylation that is preserved through evolution.

## 6 **Results**

7 *CDS and cDNAs are the less biased data and thus the best choice for a pan-species study*

8 We focused in this study on gene body DNA methylation. Annotated genomes are now available  
9 for many species, but messenger RNA sequence data is even more abundant and mRNA is  
10 representative of gene body DNA sequences. They could therefore be used instead of DNA  
11 sequence, but mRNA data is for historical reasons stored in different forms and in different  
12 databases. We reasoned that data quality will be critical for providing unbiased estimation of DNA  
13 methylation in gene bodies and therefore conducted a comparative pilot study to identify the best  
14 possible data source for the subsequent pan-species study. We used coding/transcript sequences  
15 from full genome annotations (CDS), dbEST, and cleanEST (details in Supplementary file 2). A  
16 total of 127 species are in common between CDS and dbEST, and 92 species were in common  
17 between dbEST and cleanEST. Only 29 species were common to all three databases  
18 (Supplementary file 3). We produced Notos CpGo/e profiles for all intersecting datasets and  
19 proceeded to visual inspection. In 11 out of 29 cases (38%) we identified discrepancies in at least  
20 one out of the three profiles and decided to clarify their origins by a detailed analysis of the  
21 sequences under the differential peaks. An in-depth analysis revealed that these discrepancies were  
22 either due to contaminations during the sequencing process, reflect co-occurrence of other species,  
23 or are due to bias in data acquisition. For instance, for *Trichoplax adhaerens*, *Anolis carolinensis*  
24 (green anole lizard) and *Cordyceps militaris* one or two additional shoulder peaks in dbEST and  
25 CleanEST datasets. We isolated the sequences contained in these peaks (dbEST and Clean EST)  
26 and performed a Blast2GO analysis with the aim to know their origins and functions. For the anole  
27 lizard (Supplementary figure 1), two peaks were isolated (peak 1: 0.92-1.08 and peak 2: 1.14-1.22),  
28 representing 7,030 and 4,922 sequences, respectively. The majority of sequences under peak 1 in  
29 the dbEST profile correspond to chloramphenicol O-acetyltransferase used in bacterial cloning  
30 vectors. It is therefore likely that these sequences represent contaminations from the EST library  
31 generation procedure. Sequences under peak 2 present homologies with sequences from  
32 apicomplexans (plasmodium), and platyhelminths suggesting presence of such parasites in the  
33 initial biological sample. For *T. adhaerens*, a peak was isolated (1.22-1.35), which represents 1,609

1 ESTs. Most of the sequences under the dbEST peak were identified by Blast2GO as ‘other’. Since  
2 *T. adherens* is known to contain intracellular bacteria<sup>38</sup> we believe that these sequences originate  
3 from them (Supplementary figure 2). For the mold *C. militaris*, two peaks were isolated (1.14-1.22,  
4 and 1.26-1.32). For the sequences under these peaks, homologies with other fungi sequences were  
5 found. We conclude that, in all three species, the additional modes occurred due to presence of  
6 sequences from other species, either through contamination during RNA extraction and library  
7 preparation, or as co-purification from naturally occurring symbionts or parasites.  
8 In other species, we identified other sources of bias in the transcript data. For instance, in *Bombyx*  
9 *mori* an ovarian library cleanEST showed an additional weak shoulder peak. We isolated the 769  
10 sequences under this peak (0.40-0.60). The gene ontology showed that all sequences coded for the  
11 ribosomal protein SA (RPSA). In human, RPSA genes are indeed highly expressed in the ovary but  
12 no data are available for other species. Nevertheless, it seems unlikely that the high abundance of  
13 RPSA ESTs reflects an expression bias. We speculate that the research interest of the submitters  
14 focused on this particular gene and that therefore many individual EST were submitted  
15 (Supplementary figure 3). Also in the duck (*Anas platyrhynchos*), we found a shoulder peak at 0.57  
16 and 0.59 in data from dbEST and Clean EST. Sequences under this peak corresponded to an EST  
17 library exclusively composed of immunoglobulins (146 sequences), reflecting probably a bias  
18 introduced by specific research interests (Supplementary figure 4). Interestingly, these sequences  
19 had a CpGo/e ratio between 0.5 and 0.8 suggesting hypomethylation, and in human,  
20 immunoglobulin genes in lymphoid cells are indeed undermethylated during differentiation<sup>39</sup>.

21 Finally, when we compared profiles with two peaks (bimodality), where we had noticed differences  
22 between CDS (derived from genomes) and dbEST/CleanEST (mRNA) for three invertebrate and  
23 one plant species (*Crassostrea gigas*, *Nasonia vitripennis*, *Nematostella vectensis* and *Oryza sativa*)  
24 (Supplementary figures 5-8): mRNA derived profiles showed a higher peak in genes predicted to be  
25 methylated. Gene body methylation is suspected to increase transcription<sup>40,41</sup>. The principal  
26 differences between CDS and EST data is that for the former only one FASTA sequence per gene is  
27 considered while for the later potentially several FASTA sequences for a gene could be present. We  
28 therefore hypothesized that RNA abundance induced the bias in EST data. To test this hypothesis,  
29 we performed a RNA-seq analysis in these four species. We found that genes under the low CpGo/e  
30 peak (presumably hypermethylated) show higher median RNA amounts than genes under the high  
31 CpGo/e peak (this presumably hypomethylated). mRNA FPKM medians are 1.95 to 5.45 higher in  
32 presumably hypermethylated gene bodies (Supplementary figure 5-8). We conclude that this  
33 expression difference is probably the origin of the bias in EST datasets.

1 In summary, dbEST and cleanEST have the advantage of being large repositories with data for  
2 many species, but for the purpose of our study we considered them too noisy. A complete list of  
3 species for each dataset is in supplementary file 4.

#### 4 *CpGo/e clustering identifies four types of gene body DNA methylation*

5 After having firmly established that cDNA provides an unbiased data basis, CpG o/e clustering was  
6 carried out on the 142 species for which CDS or cDNA were available. Parameters for mode  
7 number (n), mode positions (Mo), skewness (sk), and standard deviation (SD) of CpGo/e values  
8 were iteratively changed using species with known gene body DNA methylation. For further  
9 analysis, we used the following features that produced four clusters of CpGo/e: (cluster 1) species  
10 with one mode  $Mo \geq 0.69$  and  $SD < 0.12$ , (cluster 2) species with one mode  $Mo \geq 0.69$  and  $SD \geq$   
11  $0.12$ , (cluster 3) species with one mode  $Mo < 0.69$ , and (cluster 4) species with (a) two modes or (b)  
12 one mode and a skewness smaller than  $-0.04$ . Results are in supplementary file 5 and supplementary  
13 figure 9. We then associated the four clusters with known methylation types.

14 Fourteen species (9.72 %), from different phylogenetic groups (*e.g.* Ascomycota, Apicomplexa,  
15 Basidiomycota, Plathyhelminthes and Arthropoda) constitute the cluster 1. All the species have a  
16 CpGo/e mode position mode above 0.69 (the mean CpGo/e peak position is 1.00), a weak negative  
17 skewness ( $mean_{absolute\ Q50\ skewness} = -0.0019$ ) and a narrow standard deviation ( $mean_{SD} = 0.11$ ). For 4  
18 species (29%), experimental data on DNA methylation was available in the literature. All these  
19 species showed either absence of DNA methylation in the gene bodies or extremely low levels  
20 (Supplementary file 5). The latter was found in only one species (*Chlamydomonas reinhardtii*) where  
21 WGBS revealed methylation in exons but still it was 20-30 times weaker compared to other plant  
22 species in the same study. We qualify cluster 1 as “ultra-low gene body methylation” (type 1).

23 It could be argued that absence of gene body methylation is simply a consequence of absence of  
24 enzymatic methylation activity. We therefore performed a metanalysis of existing literature data  
25 concerning DNMT presence. In cluster 1, 6 out of 14 species have DNMT2 or TRDMT1, and one  
26 specie has DNMT1 and DNMT2. Only the *de-novo* methylase DNMT3 is absent in this cluster.  
27 Absence of methylation does therefore not indicate necessarily absence of DNMT genes  
28 (Supplementary file 5).

29 Cluster 2 is constituted by 60 species (41.67 %), also from different phylogenetic groups  
30 (apicomplexa, oobionta, rhodonbionta, ascomycota, basidiomycota, nematoda, platyhelminthes,  
31 annelida, arthropoda, ctenophora, chordata, embryophyta). As in the cluster 1, species present in the  
32 second cluster have a mode position  $> 0.69$  with a mean mode position very close to the first cluster



1 (mean CpGo/e position is 0.92) and a mean absolute Q50 skewness of 0.0012. However, in contrast  
2 to cluster 1, a wide standard deviation (mean<sub>SD</sub> = 0.18) has been observed. Literature data were  
3 available for 18 species (30%). For 6 species (*Schizosaccharomyces pombe*, *Aspergillus flavus*,  
4 *Brugia malayi*, *Meloidogyne incognita*, *Tribolium castaneum*, *Drosophila melanogaster*) no  
5 methylation was reported. Methylation in 6 species (*Schistosoma mansoni*, *Schistosoma japonicum*,  
6 *Fasciola hepatica*, *Petromyzon marinus*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*) is  
7 controversial since different authors come to different conclusions. Nevertheless, methylation  
8 seems to be very low. Only three species (*Trichinella spiralis*, *Solenopsis invicta*, *Physcomitrella*  
9 *patens*) showed DNA methylation in gene bodies. DNMTs were studied in 37 species. In this  
10 cluster, 16 species have just DNMT2. One species has DNMT1 or TRDMT1 only, and 6 species  
11 have all the DNMT (DNMT1, 2 and 3). Interestingly, two species (*Selaginella moellendorffii*,  
12 *Physcomitrella patens*) has just DNMT1 and 3, and two just DNMT1 and 2 (*Tribolium castaneum*,  
13 *Gasterosteus aculeatus*). Also, for *Strigamia martima* DNMT1 and DNMT3 were found, but  
14 DNMT2 was not searched for (Supplementary file 5). We consider cluster 2 as “low gene body  
15 methylation” (type 2).

16 Species with a mode position  $\leq 0.69$  form the cluster 3 (mean CpGo/e position is 0.45). The  
17 skewness is positive and larger than in the two first clusters (mean<sub>absolute Q50 skewness</sub> = 0.0483). The  
18 standard deviation (mean<sub>SD</sub> = 0.19) is wider than the cluster 1 and 2. Forty-three species (29.86 %)   
19 are present in this cluster, belonging to various phylogenetic groups (apicomplexa, sponges and  
20 nematodes, arthropoda, a large panel of chordata, and embryophytes). Many of them are important  
21 model organisms. For 26 (60%) literature data on DNA methylation was found. Gene bodies are  
22 methylated. In 19 species DNMTs were analyzed. Twelve species have all three DNMTs. One  
23 species (*Naegleria gruberi*) has DNMT1 and 2, but methylation of DNA was not yet studied. Four  
24 species have only DNMT2. Two species has DNMT1 and 3, and one only DNMT1 (Supplementary  
25 file 5). We qualify this cluster as with “gene body methylation” (type 3).

26 Finally, cluster 4 contains species that show bimodality or are strongly negatively skewed CpGo/e  
27 distributions (mean<sub>absolute Q50 skewness</sub> = -0.0424, mean CpGo/e position of mode 1 is 0.54, of mode 2  
28 0.85). Twenty-seven species (18.75 %) from different phylogenetic groups compose this cluster  
29 (apicomplexa, cnidaria, nematoda, arthropoda, mollusca, tunicata, embryophyta). Five species are  
30 strongly negatively skewed and 15 species are bimodal. We found information on DNA  
31 methylation for 10 species (50%). All species show a mosaic type of methylation with DNA regions  
32 of ultra-low methylation interspersed with regions of strong methylation. Eleven species out of 15  
33 that were studied have the three DNMT (1, 2 and 3), two had DNMT1 and 2, and two DNMT1 and

1 3 with uncertainty about DNMT2 (Supplementary file 5). Species in this cluster were considered as  
2 “mosaic type DNA methylation” (type 4).

3 Decision criteria are summarized in Figure 1.

4 *DNA methylation types do not follow the tree of life but are consistent within major clades*

5 Among unicellular organisms, kinetoplastids are unmethylated, while alveolate protists are  
6 generally methylated (Figure 2) with secondary loss of methylation that can lead to weaker  
7 methylation level or mosaic methylation. In flowering plant species, we differentiate either high  
8 (probably global) methylation in dicotyledons, and mosaic methylation in poaceae and potentially  
9 all monocotyledons. Fungi process in general ultra-low to weak methylation in the gene bodies.  
10 Also, platyhelminthes are characterized by low methylation in the coding regions. Gene bodies of  
11 deuterostomes are in general strongly methylated. There are, however, peculiar cases, *e.g.* two  
12 tunicate species (*Ciona intestinalis* and *C. savignyi*) that diverged from each other 184 ( $\pm 15$ ) Mya  
13 are in two different clusters (types 4 and 2, respectively)<sup>42</sup> with mosaic and weak methylation,  
14 probably due to secondary loss of methylation. Within lophotrochozoa, annelids show low gene  
15 body methylation, and all studied mollusks are of the mosaic type. Nematodes have in general weak  
16 gene body methylation. A particular interesting and heterogenic clade concerning methylation types  
17 are arthropods. All tested diptera (and potentially all antliophora) belong to the low methylation  
18 clusters 1 and 2, certainly due to secondary loss of methylation after splitting from its insect sister  
19 clades. All other orders show weak to high methylation with occasionally mosaic type, probably  
20 through secondary gain of local methylation.

## 21 **Discussion**

22 Evolution is based on the selection of phenotypic variants that must (i) confer a reproductive  
23 advantage to the individual, and (ii) are heritable, *i.e.* information how to generate the phenotypic  
24 variants in response to an environment are passed from parents to offspring. Heritability has  
25 traditionally been thought to be exclusively genetic, *i.e.* based on variations in the DNA sequence.  
26 In this view, genetic information is then expressed under influence of environmental cues to bring  
27 about the phenotype, a process known as  $G \times E \rightarrow P$ <sup>43</sup>. During the last 30 years it became however  
28 clear, that a substantial amount of heritable phenotypic variance can be coded by non-genetic  
29 means<sup>44</sup>. We had earlier conceptualized this view as a systems approach to inheritance, that  
30 includes genetic, epigenetic, cytoplasmic and microbial elements that are interrelated by forward  
31 and reverse interaction<sup>2</sup>. These elements interact mutually, and with the environment, to give raise  
32 to the phenotype. In this concept, genetic information (the genotype) is only one of many elements

1 that as part of an inheritance system providing heritable information, that the environment will  
2 shape into a phenotype. We define here ‘inheritance system’ as a system that is able to write and  
3 store, transmit, and receive hereditary information<sup>45</sup>. The concept implies also that genotype and  
4 epigenotype cannot exist independent of each other, and are interrelated by forward action and feed-  
5 back. This is different from the idea that sees the genome as hard-wired information that is  
6 controlled by the epigenome<sup>46</sup>. In the latter, the epigenome is conceptually closer to the (molecular)  
7 phenotype (*i.e.* product of the genotype) than to an element of the inheritance system itself.

8 The introduction of the epigenotype notion did not really solve the question, since theoretically each  
9 phenotype could just be the visible expression of its underlying epigenotype. Given the multiple  
10 facets phenotypes can acquire in living organisms, it is remarkable that, with very few exceptions,  
11 the genetic material and the genetic code remains extremely constant and thus universal<sup>47</sup>. In other  
12 words, there exists a single ‘type’ of genome. The origin of this universality of the genetic code  
13 remains enigmatic and controversial but whatever the origin is, it allows to transmit coded  
14 information from one generation to the next. These generations can understand the code since it  
15 uses a universal and constant key.

16 Given the presumably close relation between genotype and epigenotype we and others reasoned that  
17 the epigenotype and the epigenetic code should equally possess universality. The high conservation  
18 of histones and histone marks, and the conservation of methylation of cytosines suggests indeed  
19 this. Nevertheless, one could argue that the epigenetic code is simply entirely genetically  
20 determined. If this were true, we would expect that different DNA methylation types would  
21 correspond to the clades in taxonomical tree that are based on DNA sequence similarity. Our results  
22 do not support this view. Alternatively, DNA methylation types could entirely be determined by  
23 environmental conditions. In this case, similar environments should impose similar DNA  
24 methylation types. Neither our results, nor recent analyses of DNA methylation in invertebrates  
25 provide evidence of this. *E.g.* a very comprehensive study of DNA methylation in insects<sup>37</sup> did not  
26 find relations of methylation types to social behavior and the authors concluded that DNA  
27 methylation must have “more ubiquitous function”. However, compared to the tremendous amount  
28 of genomic data that is available, epigenomic data is relatively sparse and biased, which is an  
29 obstacle to answer the question conclusively. In the present study, we coped with this caveat by  
30 using a hybrid approach in which we combined available experimental data on DNA methylation  
31 with results coming from a newly developed software that predicts gene body DNA methylation  
32 types with CpG o/e ratios. Our algorithm (based on the number of species positive predicted and  
33 true positives based on the literature) allowed for including species for which no experimental DNA  
34 methylation data existed. The PPV of the algorithm is excellent for mosaic methylation (PPV=1),  
35 and methylated gene bodies (PPV=0.875), but decreases then to 0.75 (low methylated) and 0.5 for

1 ultra-low gene body methylation. This is due to the fact that our algorithm does not differentiate  
2 well between low and ultra-low methylation. If we consider the dataset as a whole, out of the 54  
3 species with known DNA methylation types, 41 were predicted correctly (total PPV = 0.76).  
4 There are some particularly interesting cases of “wrong” prediction. *Cryptosporium parvum* is a  
5 monoxenic unicellular parasite of vertebrates. It belongs to the apicomplexan its exact phylogenetic  
6 position is controversial. Exysted oocysts are the only stage that can be used to produce host DNA  
7 free genomic DNA preparations. Notos predicts clearly high gene body methylation but LC-ESI-  
8 MS did not detect 5mC in exysted oocysts purified from infected cattle<sup>48</sup>. Genome analysis of *C.*  
9 *hominis* to which *C. parvum* has only 3–5% sequence divergence<sup>49</sup>, showed that the number of  
10 genes is reduced (3,952 genes) compared to other apicomplexan, relying heavily on host gene  
11 activity. The genome shows also traces of integration of genes by lateral transfer. We hypothesize  
12 that either the progenitor DNA was methylated, or that cryptosporidium methylates DNA in the  
13 intracellular stages using the vertebrate host DNMTs.

14 Another peculiar case is the ciliate *Tetrahymena thermophile*. Also for this species Notos predicted  
15 high methylation while radioisotope labeling showed that *Tetrahymena* contains only N6-methyl-  
16 adenine but not 5mC<sup>50</sup>. *T. thermophila* and other ciliates use DNA elimination to remove  
17 approximately one-third of the genome, when the somatic macronucleus differentiates from the  
18 germline micronucleus. Histone 3 lysine 9 trimethylation (H3K9me3) is deposited on DNA  
19 destined for this elimination (reviewed in Bracht<sup>51</sup>). Interestingly, in other ciliates, DNA  
20 methylation is used for the tagging of DNA to be eliminated. It might therefore be that *Tetrahymena*  
21 had used DNA methylation in the past and has lost this capacity relatively recently, so that we still  
22 see traces in the CpG o/e ratio.

23 In summary, Notos predicts very reliable mosaic and high gene body methylation without being  
24 entirely error free. We had earlier<sup>36</sup> used only mode number (1 or 2, *i.e.* non-mosaic and mosaic  
25 methylation) and peak position of 0.75 to differentiate species with presumably methylated (<0.75)  
26 and non-methylated ( $\geq 0.75$ ) gene bodies. For the present work we added skewness -0.04, and SD  
27 0.12, and changed peak position threshold to 0.69 for better prediction.

28 Conceptually, our approach is based on the classical observation that CpN dinucleotides are  
29 observed in statistically expected frequency in low methylated regions or genomes. It was initially  
30 used to identify unmethylated CpG island in vertebrate promoters, and two major algorithms exist  
31 (Gardiner-Garden and Frommer<sup>52</sup> and Takai and Jones<sup>53</sup>). These two algorithms use the CpG o/e  
32 ratios with a score above 0.60 and 0.65, respectively. ‘Score’ (here ‘mode position’ Mo) is one  
33 parameter of our clustering algorithm. We used a decision tree to iteratively adjust this score and  
34 reached 0.69. This value is close to what was used in previous studies (*e.g.* for *C. intestinalis*: 0.70<sup>54</sup>  
35 and 0.80<sup>31</sup>, and *Nematostella vectensis*, 0.70<sup>54</sup>, or *Apis mellifera*, 1.0<sup>54</sup>). It is conceivable that Mo

1 could be slightly different for each major phylogenetical clade, and using more sophisticated  
2 clustering algorithms such as support vector machine clustering that can use multiple thresholds  
3 could still improve the PPV of our method. In addition, more experimental data on a wide range of  
4 organisms is urgently needed.

5 We find that there are four types of gene body methylation. Despite a much wider data basis in  
6 terms of phylogenetic clades, our results confirm earlier findings that concluded on three to four  
7 DNA methylation types<sup>17,24</sup>. This could be the result of a “frozen accident” situation in which  
8 methylation (e.g. type 1 and type 3) occurred randomly in early ancestors (since 5mC is coding  
9 neutral that would not have had an impact on translation), but with the establishment of a chromatin  
10 structure 5mC was recruited as epigenetic information carrier, and any change in DNA methylation  
11 type would have had a strong impact on genome function and thus fitness and was therefore  
12 maintained. Nevertheless, switching of methylation type has occurred in evolutionary time scales.  
13 Our findings indicate that there were at least three large events of secondary loss of DNA  
14 methylation: in archaeplastida (the “true” plants) where we find one branch with high methylation  
15 and another with mosaic methylation (in monocotyledons), ultra-low or mosaic methylation in the  
16 apicomplexa branch of "protists", and one transition to ultra-low gene body methylation in Diptera  
17 (Figure 2). For *D. melanogaster* in the dipteran branch it was shown experimentally that only the  
18 ‘writing’ capacity of the epigenetic inheritance element was lost, not the receiving (‘reading’)  
19 capacity<sup>55</sup>. The reason for evolutionary switching between methylation types is not clear and  
20 arguments are controversial.

21 It has been proposed that secondary loss of DNA methylation occurs because its mutational costs  
22 outweighed its adaptive value<sup>56</sup>. Indeed, in mosaic type methylation it is the evolutionary stable  
23 “old” genes that are in the methylated compartments meaning that there must be stabilizing  
24 mechanisms that prevent mutations there. Therefore, it might not be the mutational costs but the  
25 costs of maintaining such mechanisms that becomes an evolutionary burden. It was an early  
26 observation that CpG containing codons are used much less in coding sequences of vertebrates, and  
27 mutations due to CpG methylation was considered a major cause for such codon bias<sup>57</sup> and therein.  
28 Codon bias was observed also recently in the reef-building coral *Acropora millepora*<sup>57</sup>, and linked  
29 to mosaic methylation in this species. Again, phylogenetically old genes which are constitutively  
30 expressed are methylated and CpG depleted. The authors conclude that CpG methylation leads to  
31 mutations that establish a set of preferred codons in constitutively expressed genes. Once such  
32 codon bias is fixed, then alleles that control the abundance of appropriate tRNAs could have  
33 stronger effects more amenable to natural selection. The authors hypothesize that an advantage of  
34 mutation-driven codon bias that it would be beneficial for organisms with small population size or  
35 otherwise inefficient selection. Still another explanation for mosaic methylation was advanced by

1 Gavery and Roberts<sup>32</sup> who speculated that hypo-methylated regions (here in the pacific oyster  
2 *Crassostrea gigas*) might have greater epigenetic flexibility and higher regulatory control than  
3 hyper-methylated ones. Mosaic methylation could also be the result of whole genome duplication  
4 (WGD) events as suggested for *Oryza sativa*<sup>56</sup>. In addition, we have shown that environmental  
5 conditions can influence on germ-line methylation in *C. gigas* that possess mosaic methylation, and  
6 that blocks of CpG methylation are added or removed preferentially in or around genes<sup>58</sup>. One  
7 should keep in mind that DNA methylation is only one of many bearers of epigenetic information.  
8 Another one, and probably the most difficult to capture is the topology of the interphase nucleus.  
9 Using Hi-C data, Lieberman-Aiden *et al.*<sup>59</sup> established that the human genome is divided into two  
10 compartments (A-B) with pairs of loci in compartment B showing higher interaction frequency at a  
11 given genomic distance than pairs of loci in compartment A. They concluded that compartment B is  
12 more densely packed (heterochromatic) than compartment A. Higher average DNA methylation  
13 was later found to be a good predictor for the open compartment A in human cell lines<sup>60</sup> but that  
14 link could be broken in cancer cells. This cannot be interpreted as DNA methylation being decisive  
15 for topologically associated domains (TAD) establishment since DNA methylation free organisms  
16 such as *D. melanogaster* also presents canonical A-B domains<sup>61</sup>. But in drosophila, such TAD  
17 organization is not driven by long-lived interactions but rather relies on the formation of transient,  
18 low-frequency contacts<sup>62</sup>. We hypothesize therefore that DNA methylation actually impacts on the  
19 relative dynamics of formation of contacts in A and B compartments, possibly stabilizing them. It is  
20 tempting to speculate that one consequence of compartmentation of genomes dynamics by  
21 methylation is that this might create additional units of selection. Results from tunicates support this  
22 idea: *Ciona* CpG<sub>0/e</sub> ratios have different profiles (bimodal for *C. intestinalis* and unimodal for *C.*  
23 *savignyi*). The *C. intestinalis* methylome is predicted to be mosaic that corresponds to experimental  
24 observations<sup>63</sup>. Our prediction for *C. savignyi* is low methylation (cluster 2). Both species diverged  
25 from each other 184 ( $\pm 15$ ) Mya<sup>42</sup> and their genomes are very different<sup>64,65</sup>. For instance, analysis of  
26 18S rRNA sequences shows that the pairwise divergence of the two *ciona* species is slightly greater  
27 than that between human and *e.g.* birds<sup>66</sup>. This is puzzling since developmental features, body plan,  
28 effective population size and environment are very similar, and even hybrids can be generated to  
29 the tadpole stage<sup>67</sup>. However, *C. savignyi* shows a genome wide average Single Nucleotide  
30 Polymorphism (SNP) heterozygosity of 4.5% while *C. intestinalis*, that has mosaic methylation, is  
31 genetically less polymorphic (1.5%) (reviewed in Veeman *et al.*<sup>68</sup>). It is conceivable that the  
32 methylated *C. intestinalis* genome can generate sufficiently stable TADs so that genome x  
33 epigenome interactions can serve as heritable unit of selection, while in *C. savignyi* TADs are more  
34 dynamic because the relative weight of DNA methylation in the generation of stable heritable  
35 phenotypic variants is less important. Our prediction concords with very recent results showing that

1 stress-induced DNA methylation changes in *C. savignyi* can occur but are highly ephemeral (<48-  
2 120 h), and thus not maintained through germline<sup>69</sup>.  
3 In conclusion, our findings indicate that initially there were three types of gene body DNA  
4 methylation: ‘primary no methylation’, ‘primary whole genome methylation’, and ‘primary mosaic  
5 methylation’ that produced by secondary loss ‘weak methylation’, or ‘secondary no methylation’.  
6 These findings are in concordance with the idea that DNA methylation in gene bodies (i) uses three  
7 types of universal codes (low, high and mosaic)), and (ii) that it is an element of the inheritance  
8 system and not a molecular phenotype that results from genotype x environment interaction. This  
9 has immediate practical consequences: e.g. since there are three types of methylation codes, pan-  
10 species conclusions about the potential function of DNA methylation can only be drawn within the  
11 type (e.g. functional tests in vertebrates with high gene body methylation cannot be used to  
12 conclude on methylation function in mosaic type mollusks). In addition, if DNA methylation is part  
13 of the inheritance system then heritable phenotypic diversity can be produced by DNA methylation  
14 changes without changes in the DNA sequence. The notion that everything that is heritable is  
15 necessarily genetic should be abandoned.

16

## 17 **Methods**

### 18 *Origin of sequences, data cleaning and Notos parameters*

19 In this study, coding sequences (CDS) or cDNA sequences of 147 species were downloaded from  
20 Ensembl and VEGA databases. Expressed sequence tags (ESTs) were downloaded from two  
21 different databases: dbEST<sup>70</sup> (605 species) and CleanEST<sup>71</sup> (110 species) (Supplementary file 3).  
22 We used Notos<sup>36</sup> for the calculation and modelling of CpGo/e distribution with the three datasets,  
23 with a minimal length L=200 bp and formula 1<sup>72</sup>:

24

[Formula 1]

$$CpG\ o/e = \frac{\text{number of CpG}}{\text{number of C} \times \text{number of G}} \times \frac{L^2}{L - 1}$$

25 All the values outside the interval, and all the values with a score of 0 were removed. For each  
26 species, the number of mode, the position of mode(s), the Q50 skewness coefficient and the  
27 standard deviation (SD) were calculated.

### 28 *Blast searches and gene ontology analysis*

1 Database searches were done by Blastx searches against a local instance (ncbi-blast-2.2.30+) of  
2 non-redundant ‘nr’ with 20 maximum hits, an E-value of 0.001, and other parameters as default  
3 values. Gene ontology searches were performed with blast2go<sup>73</sup>.

#### 4 *RNA seq analysis*

5 RNAseq datasets for *Nematostella vectensis*, *Nasonia vitripennis*, *Crassostrea gigas* and *Oryza*  
6 *sativa japonica* were downloaded as fastq files from the European Nucleotide Archive and NCBI  
7 (details in supplementary file 3). For each dataset, the reads were filtered with a Fred quality score  
8  $\geq 26$ . Filtered reads were mapped on their reference genomes (downloaded from Ensembl, details in  
9 supplementary file 3) with RNA STAR<sup>74</sup> on a local Galaxy instance (v2.4.0d-2). Resulting BAM  
10 files and the gff files (downloaded from Ensembl, details in supplementary file 3) with the coding  
11 sequences were used for FPKM estimations with Cufflinks<sup>75</sup>. Annotation gff-files were used to  
12 extract CDS in fasta format from their genomes and we calculated the CpGo/e ratios with Notos<sup>36</sup>  
13 and detected modes (peaks). To compare FPKM for the genes under the peaks, a bandwidth of 0.2  
14 ( $\pm 0.1$  around mode maximum) was arbitrarily chosen for the CpG o/e ratio. FPKMs were extracted  
15 and used for statistical analysis of expression level in gene bodies with low and high predicted  
16 methylation. Mood’s median test was used with R<sup>76</sup>.

#### 17 *Meta-analysis of DNA methylation using literature data*

18 For each species for which data was available in the above-mentioned databases, we searched the  
19 literature on Google scholar (as of April-June 2016) with the following keywords: DNA  
20 methylation, 5-methyl-cytosine, gene body, mosaic methylation, global methylation, DNA  
21 methylation pattern. Articles were obtained from Bib CNRS (<https://bib.cnrs.fr/>) and manually  
22 curated to obtain gene body methylation, and presence of DNMTs.

#### 23 *Clustering*

24 To identify distinct subgroups within the 147 analyzed species, we generated descriptive analyses,  
25 considering both the KDE of the CpGo/e ratios and aggregating statistics based on it. The statistics  
26 we used were (1) the number of modes of the KDE, (2) the position of the modes, (3) the standard  
27 deviation SD of the CpGo/e ratios, (4) absolute Q50 mode skewness of the CpGo/e ratios, *i.e.* ,

28 [Formula 2]

$$\frac{Q_3 + Q_1}{2} - Mo$$



1 with  $Q_1$  and  $Q_3$  the 25 % and 75 % quantile of the CpGo/e ratios, respectively, and Mo the global  
2 mode of the KDE. We investigated several formulas for the skewness, deeming the absolute Q50  
3 mode skewness the most informative for our analysis<sup>36</sup>. For the sake of readability, we refer to the  
4 absolute Q50 mode skewness as “skewness” in what follows.

5 The four clusters into which we classify the species are specified in the result section. The values of  
6 the three thresholds used in the definition of the clusters were determined by evaluating the  
7 prediction performance of our approach depending on these three values, using 54 species for which  
8 the true methylation type had been determined experimentally (given in Supplementary file 5).  
9 Hereby, the clusters correspond to the patterns used in that file like follows: Cluster 1 - not  
10 methylated; Cluster 2 - low methylated; Cluster 3 - (high) methylated / global methylation; Cluster  
11 4 - mosaic methylation.

12 To determine the optimal threshold values, we employed a two-step approach. First, we searched  
13 the whole parameter space (*i.e.* the space of all possible values the thresholds can assume) using a  
14 Metropolis-Hastings algorithm to ensure that strongly deviating from the threshold values we chose  
15 manually always leads to a poor prediction performance. Second, we systematically searched the  
16 parameter space around the manually chosen values. That is, we evaluated the prediction  
17 performance on a grid of size  $21^3 = 9261$ , covering the following threshold values: skewness from -  
18 0.08 to -0.04 in steps of 0.002; peak position from 0.69 to 0.79 in steps of 0.005; SD from 0.11 to  
19 0.21 in steps of 0.005. For the present work we used: skewness -0.04, peak position 0.69, and SD  
20 0.12.

21 Due to the scarcity of the data, the optimal prediction (76 %, 41 out of 54 true) is achieved for a  
22 rather large set of threshold values. To judge the performance of our algorithm, it should be noted  
23 that for 7 out the 13 misclassified species the true and the predicted classifications are “not  
24 methylated” and “low methylated”, or *vice versa* respectively. That is, the mistake made by the  
25 algorithm is rather small in these cases. The remaining four wrong predictions are actually peculiar  
26 cases that were discussed above.

27 The clustering was implemented using R version 3.4.0 (supplementary files 6 and 7). For  
28 visualizing the clustering, the R package `dendextend` has been used. Parameters were Rscript  
29 `cluster.r -0.04 0.69 0.12 input_file_notos input_file_notos_bootstrap` with first parameter  
30 [skewness], second [Mo], and third [SD], and Notos outputfiles as input. Further details on our  
31 method can be found in<sup>25</sup>.

32 *Tree of life*

1 We recovered the taxonomic IDs of all investigated species from the NCBI taxonomy database  
2 (<https://www.ncbi.nlm.nih.gov/taxonomy>) and created a tabular file (.txt). We used this file to  
3 generate a Phylip tree file based on the classification in the NCBI taxonomy database with the  
4 NCBI common taxonomy tree online tool  
5 (<https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi>) and designed a tree with the  
6 interactive Tree of life (version 4.0.2) (<https://itol.embl.de>)<sup>77</sup>.

## 7 References

- 8 1. Levine, A. J. The Future of Systems Biology. *Curr. Opin. Syst. Biol.* **1**, v–vii (2017).
- 9 2. Cosseau, C. *et al.* (Epi)genetic Inheritance in *Schistosoma mansoni*: A Systems Approach. *Trends*  
10 *Parasitol.* **33**, 285–294 (2017).
- 11 3. Nicoglou, A. & Merlin, F. Epigenetics: A way to bridge the gap between biological fields. *Stud. Hist.*  
12 *Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* 1–10 (2017).  
13 doi:10.1016/j.shpsc.2017.10.002
- 14 4. Hotchkiss, R. D. The quantitative separation of purines, pyrimidines and nucleosides by paper  
15 chromatography. *J. Biol. Chem.* **175**, 315–332 (1948).
- 16 5. Ye, P. *et al.* MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine  
17 generated by single-molecular real-time sequencing. *Nucleic Acids Res.* **45**, D85–D89 (2017).
- 18 6. Chen, W., Yang, H., Feng, P., Ding, H. & Lin, H. iDNA4mC: identifying DNA N4-methylcytosine  
19 sites based on nucleotide chemical properties. *Bioinformatics* **33**, 3518–3523 (2017).
- 20 7. Vanyushin, B. F. in *DNA Methylation: Basic Mechanisms* 67–122 (Springer-Verlag, 2006).  
21 doi:10.1007/3-540-31390-7\_4
- 22 8. Cambareri, E., Jensen, B., Schabtach, E. & Selker, E. Repeat-induced G-C to A-T mutations in  
23 *Neurospora*. *Science (80- )*. **244**, 1571–1575 (1989).
- 24 9. Lyko, F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev.*  
25 *Genet.* (2017). doi:10.1038/nrg.2017.80
- 26 10. Riggs, A. D., Xiong, Z., Wang, L. & LeBon, J. M. Methylation dynamics, epigenetic fidelity and X  
27 chromosome structure. *Epigenetics* **793**, 214 (2008).
- 28 11. Hermann, A., Schmitt, S. & Jeltsch, A. The human Dnmt2 has residual DNA-(Cytosine-C5)  
29 methyltransferase activity. *J. Biol. Chem.* **278**, 31717–31721 (2003).
- 30 12. Goll, M. G. Methylation of tRNA<sup>Asp</sup> by the DNA Methyltransferase Homolog Dnmt2. *Science (80-.*  
31 *)*. **311**, 395–398 (2006).
- 32 13. Albalat, R. Evolution of DNA-methylation machinery: DNA methyltransferases and methyl-DNA  
33 binding proteins in the amphioxus *Branchiostoma floridae*. *Dev. Genes Evol.* **218**, 691–701 (2008).
- 34 14. Schaefer, M. & Lyko, F. Solving the Dnmt2 enigma. *Chromosoma* **119**, 35–40 (2010).
- 35 15. Raddatz, G. *et al.* Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc. Natl.*  
36 *Acad. Sci.* **110**, 8627–8631 (2013).
- 37 16. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA  
38 (cytosine-5) methyltransferases. *Nat. Am. Inc.* **19**, 219–220 (1998).
- 39 17. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat.*  
40 *Rev. Genet.* **9**, 465–76 (2008).
- 41 18. Rivenbark, A. G. *et al.* Epigenetic reprogramming of cancer cells via targeted DNA methylation.  
42 *Epigenetics* **7**, 350–360 (2012).
- 43 19. Casimir, C. M., Gates, P. B., Patient, R. K. & Brockes, J. P. Evidence for dedifferentiation and  
44 metaplasia in amphibian limb regeneration from inheritance of DNA methylation. *Development* **104**,  
45 657 LP-668 (1988).
- 46 20. Mugatroyd, C., Wu, Y., Bockmühl, Y. & Spengler, D. The janus face of DNA methylation in aging.  
47 *Aging (Albany. NY)*. **2**, 107–110 (2010).
- 48 21. Zampieri, M. *et al.* Reconfiguration of DNA methylation in aging. *Mech. Ageing Dev.* **151**, 60–70  
49 (2015).
- 50 22. Downen, R. H. *et al.* Widespread dynamic DNA methylation in response to biotic stress. *Proc. Natl.*  
51 *Acad. Sci. U. S. A.* **109**, E2183-91 (2012).

- 1 23. Cortijo, S. *et al.* Mapping the Epigenetic Basis of Complex Traits. *Science* (80-. ). **343**, 1145 LP-1148  
2 (2014).
- 3 24. Yi, S. V. & Goodisman, M. a D. Computational approaches for understanding the evolution of DNA  
4 methylation in animals. *Epigenetics* **4**, 551–556 (2009).
- 5 25. Bulla, I. *et al.* Notos - a galaxy tool to analyze CpN observed expected ratios for inferring DNA  
6 methylation types. *BMC Bioinformatics* **19**, 105 (2018).
- 7 26. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–  
8 1504 (1980).
- 9 27. Fryxell, K. J. & Moon, W. J. CpG mutation rates in the human genome are highly dependent on local  
10 GC content. *Mol. Biol. Evol.* **22**, 650–658 (2005).
- 11 28. Cooper, D. N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate  
12 genomes. *Hum. Genet.* **83**, 181–188 (1989).
- 13 29. Jabbari, K. & Bernardi, G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*  
14 **333**, 143–149 (2004).
- 15 30. Razin, A. & Cedar, H. Distribution of 5-methylcytosine in chromatin. *Proc. Natl. Acad. Sci. U. S. A.*  
16 **74**, 2725–2728 (1977).
- 17 31. Suzuki, M. M., Kerr, A. R. W., De Sousa, D. & Bird, A. CpG methylation is targeted to transcription  
18 units in an invertebrate genome. *Genome Res.* **17**, 625–31 (2007).
- 19 32. Gavery, M. R. & Roberts, S. B. DNA methylation patterns provide insight into epigenetic regulation  
20 in the Pacific oyster (*Crassostrea gigas*). *BMC Genomics* **11**, 483 (2010).
- 21 33. Park, J. *et al.* Comparative analyses of DNA methylation and sequence evolution using *Nasonia*  
22 genomes. *Mol. Biol. Evol.* **28**, 3345–3354 (2011).
- 23 34. Dixon, G. B., Bay, L. K. & Matz, M. V. Bimodal signatures of germline methylation are linked with  
24 gene expression plasticity in the coral *Acropora millepora*. *BMC Genomics* **15**, 1109 (2014).
- 25 35. Walsh, T. K. *et al.* A functional DNA methylation system in the pea aphid, *Acyrtosiphon pisum*.  
26 *Insect Mol. Biol.* **19**, 215–228 (2010).
- 27 36. Bulla, I. *et al.* Notos - a Galaxy tool to analyze CpN observed expected ratios for inferring DNA  
28 methylation types. *bioRxiv* (2017). doi:10.1101/180463
- 29 37. Bewick, A. J., Vogel, K. J., Moore, A. J. & Schmitz, R. J. Evolution of DNA Methylation across  
30 Insects. *Mol. Biol. Evol.* **34**, msw264 (2016).
- 31 38. Driscoll, T., Gillespie, J. J., Nordberg, E. K., Azad, A. F. & Sobral, B. W. Bacterial DNA sifted from  
32 the *Trichoplax adhaerens* (Animalia: Placozoa) genome project reveals a putative rickettsial  
33 endosymbiont. *Genome Biol. Evol.* **5**, 621–645 (2013).
- 34 39. Storb, U. & Arp, B. Methylation patterns of immunoglobulin genes in lymphoid cells: correlation of  
35 expression and differentiation with undermethylation. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 6642–6646  
36 (1983).
- 37 40. Bewick, A. J. & Schmitz, R. J. Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.* **36**,  
38 103–110 (2017).
- 39 41. He, X.-J., Chen, T. & Zhu, J.-K. Regulation and function of DNA methylation in plants and animals.  
40 *Cell Res.* **21**, 442–465 (2011).
- 41 42. D’Onofrio, G., Berná, L. & Alvarez-Valin, F. How fast is the sessile *Ciona*? *Comp. Funct. Genomics*  
42 **2009**, (2009).
- 43 43. Bowman, J. Genotype × environment interactions. *Genet. Sel. Evol.* **4**, 117 (1972).
- 44 44. Danchin, E. *et al.* Beyond DNA: integrating inclusive inheritance into an extended theory of  
45 evolution. *Nat. Rev. Genet.* **12**, 475–486 (2011).
- 46 45. Lamm, E. in *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) (Metaphysics Research  
47 Lab, Stanford University, 2014). at <[https://plato.stanford.edu/archives/win2014/entries/inheritance-](https://plato.stanford.edu/archives/win2014/entries/inheritance-systems)  
48 [systems](https://plato.stanford.edu/archives/win2014/entries/inheritance-systems)>
- 49 46. Laland, K. *et al.* Does evolutionary theory need a rethink? *Nature* **514**, 161–164 (2014).
- 50 47. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: The universal enigma.  
51 *IUBMB Life* **61**, 99–111 (2009).
- 52 48. Gissot, M., Choi, S. W., Thompson, R. F., Greally, J. M. & Kim, K. *Toxoplasma gondii* and  
53 *Cryptosporidium parvum* lack detectable DNA cytosine methylation. *Eukaryot. Cell* **7**, 537–540  
54 (2008).
- 55 49. Xu, P. *et al.* The genome of *Cryptosporidium hominis*. *Nature* (2004). doi:10.1038/nature02977
- 56 50. HATTMAN, S., KENNY, C., BERGER, L. & PRATT, K. Comparative study of DNA methylation in  
57 three unicellular eucaryotes. *J. Bacteriol.* **135**, 1156–1157 (1978).

- 1 51. Bracht, J. R. Beyond transcriptional silencing: Is methylcytosine a widely conserved eukaryotic DNA  
2 elimination mechanism? *BioEssays* (2014). doi:10.1002/bies.201300123
- 3 52. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282  
4 (1987).
- 5 53. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22.  
6 *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3740–5 (2002).
- 7 54. Nanty, L. *et al.* Comparative methylomics reveals gene-body H3K36me3 in *Drosophila* predicts  
8 DNA methylation and CpG landscapes in other invertebrates. *Genome Res.* **21**, 1841–1850 (2011).
- 9 55. Lyko, F. *et al.* Mammalian (cytosine-5) methyltransferases cause genomic DNA methylation and  
10 lethality in *Drosophila*. *Nat. Genet.* **23**, 363–366 (1999).
- 11 56. Zemach, A., Mcdaniel, I., Silva, P. & Zilberman, D. Genome-Wide Evolutionary Analysis of  
12 Eukaryotic DNA Methylation. *Sci. (New York, NY)* **11928**, science.1186366v1 (2010).
- 13 57. Dixon, G. B., Bay, L. K. & Matz, M. V. Evolutionary Consequences of DNA Methylation in a Basal  
14 Metazoan. *Mol. Biol. Evol.* **33**, 2285–2293 (2016).
- 15 58. Rondon, R. *et al.* Effects of a parental exposure to diuron on Pacific oyster spat methylome. *Environ.*  
16 *Epigenetics* **3**, 1–13 (2017).
- 17 59. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding  
18 Principles of the Human Genome. *Science (80-. )*. **326**, 289–293 (2009).
- 19 60. Fortin, J.-P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-  
20 range correlations in epigenetic data. *Genome Biol.* **16**, 180 (2015).
- 21 61. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila*  
22 genome. *Cell* **148**, 458–472 (2012).
- 23 62. Cattoni, D. I. I. *et al.* Single-cell absolute contact probability detection reveals that chromosomes are  
24 organized by multiple, low-frequency yet specific interactions. *Doi.Org* 159814 (2017).  
25 doi:10.1101/159814
- 26 63. Suzuki, M. M. *et al.* Identical sets of methylated and nonmethylated genes in *Ciona intestinalis* sperm  
27 and muscle cells. *Epigenetics Chromatin* **6**, 38 (2013).
- 28 64. Small, K. S., Brudno, M., Hill, M. M. & Sidow, A. Extreme genomic variation in a natural  
29 population. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5698–703 (2007).
- 30 65. Kourakis, M. J. & Smith, W. C. An organismal perspective on *C. intestinalis* development, origins  
31 and diversification. *Elife* **317**, (2015).
- 32 66. Johnson, D. S., Davidson, B., Brown, C. D., Smith, W. C. & Sidow, A. Noncoding regulatory  
33 sequences of *Ciona* exhibit strong correspondence between evolutionary constraint and functional  
34 importance. *Genome Res.* **14**, 2448–2456 (2004).
- 35 67. Byrd, J. & Lambert, C. C. Mechanism of the block to hybridization and selfing between the sympatric  
36 ascidians *Ciona intestinalis* and *Ciona savignyi*. *Mol. Reprod. Dev.* **55**, 109–116 (2000).
- 37 68. Veeman, M. T., Chiba, S. & Smith, W. C. in *Vertebrate Embryogenesis: Embryological, Cellular,*  
38 *and Genetic Methods* (ed. Pelegri, F. J.) 401–422 (Humana Press, 2011). doi:10.1007/978-1-61779-  
39 210-6\_15
- 40 69. Huang, X. *et al.* Rapid response to changing environments during biological invasions: DNA  
41 methylation perspectives. *Mol. Ecol.* **12**, 3218–3221 (2017).
- 42 70. Boguski, M. S. & Tolstoshev, T. M. J. L. C. M. dbEST-database for ‘expressed sequence tags’. *Nat.*  
43 *Genet.* **4**, 332–333 (1993).
- 44 71. Lee, B. & Shin, G. CleanEST: A database of cleansed EST libraries. *Nucleic Acids Res.* **37**, 686–689  
45 (2009).
- 46 72. Matsuo, K., Clay, O., Takahashi, T., Silke, J. & Schaffner, W. Evidence for erosion of mouse CpG  
47 islands during mammalian evolution. *Somat. Cell Mol. Genet.* **19**, 543–555 (1993).
- 48 73. Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in functional  
49 genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
- 50 74. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- 51 75. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated  
52 transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
- 53 76. R Development Core Team. R: A Language and Environment for Statistical Computing. (2008). at  
54 <<http://www.r-project.org>>
- 55 77. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation  
56 of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
- 57 78. Lyko, F. *et al.* The honey bee epigenomes: Differential methylation of brain DNA in queens and

- workers. *PLoS Biol.* **8**, (2010).
79. Fneich, S. *et al.* 5-methyl-cytosine and 5-hydroxy-methyl-cytosine in the genome of *Biomphalaria glabrata*, a snail intermediate host of *Schistosoma mansoni*. *Parasit. Vectors* **6**, 167 (2013).
80. Wurm, Y. *et al.* The genome of the fire ant *Solenopsis invicta*. *Proc Natl Acad Sci U S A* **108**, 5679–5684 (2011).
81. Simola, D. F. *et al.* Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res.* **23**, 1235–1247 (2013).
82. Wang, X. *et al.* Function and Evolution of DNA Methylation in *Nasonia vitripennis*. *PLoS Genet.* **9**, (2013).
83. Robinson, K. L., Tohidi-Esfahani, D., Lo, N., Simpson, S. J. & Sword, G. A. Evidence for widespread genomic methylation in the migratory locust, *Locusta migratoria* (orthoptera: Acrididae). *PLoS One* **6**, (2011).
84. Xiang, H. *et al.* Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat. Biotechnol.* **28**, 516-U181 (2010).
85. Cunningham, C. B. *et al.* The Genome and Methylome of a Beetle with Complex Social Behavior, *Nicrophorus vespilloides* (Coleoptera: Silphidae). *Genome Biol. Evol.* **7**, 3383–96 (2015).
86. Simmen, M. W. *et al.* Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* **283**, 1164–1167 (1999).
87. Chen, F.-C., Chuang, T.-J., Lin, H.-Y. & Hsu, M.-K. The evolution of the coding exome of the *Arabidopsis* species - the influences of DNA methylation, relative exon position, and exon length. *BMC Evol. Biol.* **14**, 145 (2014).

## Acknowledgements

The work on this tool was initiated during a meeting that had received funding of the French-Norwegian travel program AURORA. This work has been supported by Campus France and the Norges forskningsrad (program AURORA, nr. 34040YK) to C. Grunau and J. Bulla, the grant Felleslegat til fordel for biologisk forskning ved Universitetet i Bergen to J. Bulla, the ANR grant ANR-10-BLAN-1720 (EpiGEvol) to C. Grunau, a PhD grant for disabled students by the French Ministry of Education and Research to B. Aliaga, and a DFG return grant to I. Bulla (BU 2685/4-1). The authors are grateful to Céline Cosseau and Cristian Chaparro for helpful discussion.

## Author Contributions Statement

CG and DD designed the study. BA performed the experiments. IB developed the mathematical models, wrote the associated code, and did the clustering. GM and BA worked on the phylogeny. All authors wrote and reviewed the manuscript.

## Additional information

### Competing interests

The author(s) declare no competing interests.

## Figure legends

**Figure 1: Summary of decision grid for clustering of CpG o/e ratio distributions on species level.**

**Figure 2: Schematic representation of the “Tree of Life” for 147 species, associated with the four different types of DNA methylation that were identified in this work.** Numbers in brackets indicate DNA methylation types (“clusters”) for each species. Line colors correspond to methylation types.

## Tables

1 Table 1: List of publications in which the authors investigated DNA methylation by a wet bench  
 2 method and compared the results to CpGo/e ratios

Species	Formula	Sequences	Validation	References
<i>Acropora millepora</i>	Unknown	CDS	MBD-eq	57
<i>Apis mellifera</i>	Unknown	CDS	BS-seq	78
<i>Biomphalaria glabrata</i>	Matsuo <sup>72</sup>	RNAseq	Restriction enzyme, BS-seq (Nimbus retrotransposon), LC-MS	79
<i>Crassostrea gigas</i>	Matsuo <sup>72</sup>	EST	Methylation sensitive PCR, bisulfite sequencing PCR	32
<i>Solenopsis invicta</i>	Unknown	Genome	MeDIP, Bisulfite sequencing (9 genes)	80
	Gardiner-Garden and Frommer <sup>52</sup>	Promoteur and Genes	Bisulfite sequencing	81
<i>Nasonia vitripennis</i>	Matsuo <sup>72</sup>	Refseq	Cloning and sequencing 18 genes at selected CpG sites, BS-seq	33
	Unknown	Genome and coding sequences	WGBS	82
<i>Locusta migratoria</i>	Unknown	cDNA, Unigene	Methylation-specific restriction enzyme assays	83
<i>Acyrtosiphon pisum</i>	Unknown	CDS and predicted genes	MeDIP, BS-seq, restriction enzyme	35
<i>Bombyx mori</i>	Unknown	Genes	MethylC-seq	84
<i>Nicrophorus vespilloides</i>	Unknown	Genes	Whole genome bisulfite sequencing	85
<i>Ciona intestinalis</i>	Unknown	Genes	BS-Seq	86
	Unknown	EST	Bisulfite sequencing, Methylation-sensitive PCR	31
<i>Arabidopsis thaliana</i>	Gardiner-Garden and Frommer <sup>52</sup>	CDS	BS-seq	87

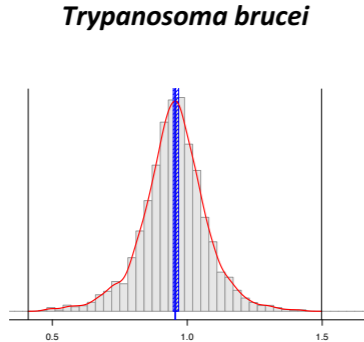
1  
2

**Cluster profile description**

**Notos parameters**

**Predicted gene body methylation type**

Cluster 1

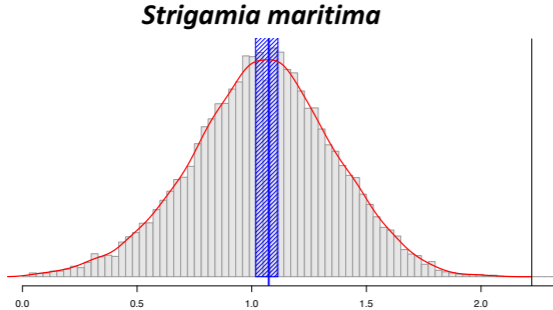


Species with a single mode, weak asymmetry and a low standard deviation.

1 mode with CpG o/e  $\geq 0.69$  and SD  $< 0.12$

Ultra-low gene body methylation

Cluster 2

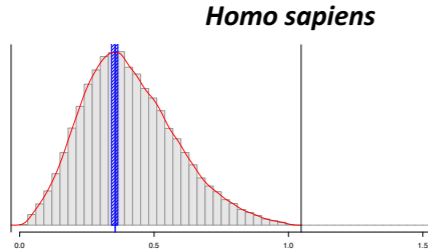


Species with a single mode, weak asymmetry and a large standard deviation.

1 mode with CpG o/e  $\geq 0.69$  and SD  $\geq 0.12$

Low gene body methylation

Cluster 3

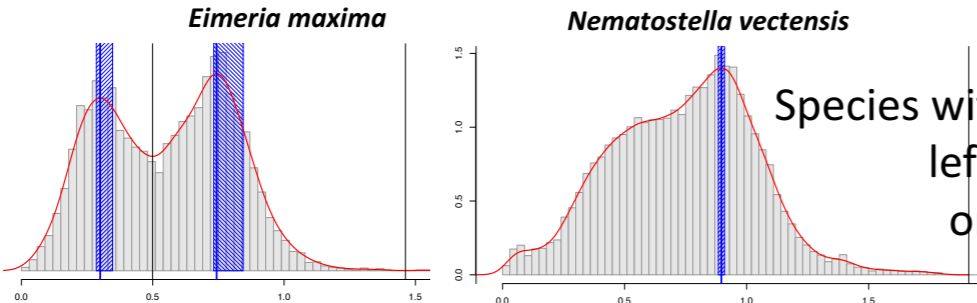


Species with a single mode and CpGo/e peak position below 0.69.

1 mode with CpG o/e  $< 0.69$

Gene body methylation

Cluster 4



Species with a single mode and left asymmetry, or bimodality

2 modes OR 1 mode and skewness  $< -0.04$

Mosaic DNA methylation type



